Received: 4 December 2019 | Accepted: 4 June 2021

DOI: 10.1111/rssa.12734

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda*

Naoki Egami[1] | **Erin Hartman**[2]

[1]Department of Political Science, Columbia University, New York, NY, USA

[2]Department of Political Science, University of California, Berkeley, Berkeley, CA, USA

Correspondence
Erin Hartman, Department of Political Science, University of California, Berkeley, Berkeley, CA 94720, USA.
Email: ekhartman@berkeley.edu, www.erinhartman.com

Abstract

Generalizing estimates of causal effects from an experiment to a target population is of interest to scientists. However, researchers are usually constrained by available covariate information. Analysts can often collect many fewer variables from population samples than from experimental samples, which has limited applicability of existing approaches that assume rich covariate data from both experimental and population samples. In this article, we examine how to select covariates necessary for generalizing experimental results under such data constraints. In our concrete context of a large-scale development program in Uganda, although more than 40 pre-treatment covariates are available in the experiment, only 8 of them were also measured in a target population. We propose a method to estimate a *separating* set—a set of variables affecting both the sampling mechanism and treatment effect heterogeneity—and show that the population average treatment effect (PATE) can be identified by adjusting for estimated separating sets. Our algorithm only requires a rich set of covariates in the experimental data, not in the target

*We thank Christopher Blattman, Nathan Fiala, and Sebastian Martinez for sharing their data. We are also grateful to Alexander Coppock, Don Green, Chad Hazlett, Zhichao Jiang, and Soichiro Yamauchi as well as the participants of the MPSA, the UCLA IDSS workshop and the Yale Quantitative Methods Workshop, for their helpful comments on an earlier version of the paper. We also thank the editor and two anonymous reviewers for providing us with valuable comments. Erin Hartman gratefully acknowledges research support from the UCLA California Center for Population Research Junior Faculty Course Release Program.

population, by incorporating researcher-specific constraints on what variables are measured in the population data. Analysing the development experiment in Uganda, we show that the proposed algorithm can allow for the PATE estimation in situations where conventional methods fail due to data requirements.

KEYWORDS

causal inference, external validity, generalization, randomized experiments

1 | INTRODUCTION

Over the last few decades, social and biomedical scientists have developed and applied an array of statistical tools to make valid causal inferences (Imbens & Rubin, 2015). In particular, randomized experiments have become the mainstay for estimating causal effects. Although many scholars agree upon the high internal validity of experimental results, there is a debate about how scientists should infer the impact of policies and interventions on broader populations (Angrist & Pischke, 2010; Bareinboim & Pearl, 2016; Deaton & Cartwright, 2018; Imai et al., 2008; Imbens, 2010). This issue of generalizability (Stuart et al., 2011) is pervasive in practice because randomized controlled trials are often conducted on non-representative samples (Allcott, 2015; Druckman et al., 2011; Egami & Hartman, 2020; Shadish et al., 2002; Stuart et al., 2015).

In this paper, we examine how to generalize the experimental results of the Youth Opportunities Program (YOP) in Uganda, which aims to help the poor and unemployed become self-employed artisans and increase incomes. This large-scale development program, involving more than 10,000 individuals, was implemented by the government of Uganda and the authors of Blattman et al. (2013) from 2008 to 2012. Young adults in Northern Uganda were invited to form groups and submit grant proposals for vocational training and to start independent trades. To evaluate the causal impact of the program, funding was randomly assigned and a host of economic variables (e.g. employment and income) were measured.

The question of generalizability is especially important in this application. The aim of such development programs is elegantly noted in Duflo and Kremer (2005), 'the benefits of knowing which programs work and which do not extend far beyond any program or agency, and credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations (NGOs) beyond national borders'. Researchers and policy makers are not just concerned to learn about the very individuals who participated in the trial. The ultimate goal is to learn whether and how much the program can improve economic conditions in a larger target population—about 10 million people in Northern Uganda (Government of Uganda, 2007).

Despite its importance, estimating population average treatment effects (PATEs) is not straightforward because we have to adjust for differences between experimental samples and the target population. One pervasive question is what covariates should and can we adjust for? Although previous research shows that adjusting for a set of variables explaining the sampling mechanism or treatment heterogeneity is sufficient for generalization (Bareinboim et al., 2014; Stuart et al., 2011), researchers are often constrained by available covariate information in applied settings.

In this paper, we address this problem of covariate selection for estimating PATEs. In particular, we develop a data-driven method to estimate *a separating set*—a set of variables affecting both the sampling mechanism and treatment effect heterogeneity. Recent papers show that the PATE can be identified by adjusting for this separating set (Cole & Stuart, 2010; Kern et al., 2016; Pearl & Bareinboim, 2014; Tipton, 2013). In Section 3, we extend this result and show that the separating set relaxes data requirements of conventional methods by generalizing two widely used covariate selection approaches: (1) *a sampling set*—a set of variables explaining how units are sampled into a given experiment (Buchanan et al., 2018; Hartman et al., 2015; Pressler & Kaizar, 2013) and (2) *a heterogeneity set*: a set of variables explaining treatment effect heterogeneity (Kern et al., 2016; Nguyen et al., 2017).

In Section 4, we demonstrate that such separating sets are estimable from the experimental data and provide a new estimation algorithm based on Markov random fields (MRFs). This algorithm only requires that a sampling set be observed in the experimental sample, not in the target population. We estimate a separating set as a set that makes a sampling set conditionally independent of observed outcomes in the experimental data. Therefore, in contrast to conventional methods, we can exploit all covariates in the experiment to find necessary separating sets, even when there are few variables measured in both the experimental and population data.

Importantly, our proposed approach maintains a widely used assumption that a sampling set is observed in the experimental data. However, unlike many existing methods, we do not assume that a sampling set is also observed in the population data. This distinction in data requirement is subtle and yet practically essential because in many applied contexts, a larger number of covariates are measured in the experimental data than in the population data. For example, the experimental data of Blattman et al. (2013) contain about 40 pre-treatment covariates, even though only 8 of them are also measured in the target population. To estimate separating sets, our proposed method incorporates such user-constraints on what variables can feasibly be collected in the target population. For instance, suppose people selected into the YOP due to social connections, which were unmeasured in the target population. Even in this scenario where conventional methods fail, the proposed method can estimate separating sets accounting for this data constraint, if any exist.

Our article builds on a growing literature on the PATE, which has two general directions. First, many previous studies have focused on articulating identification assumptions and proposing consistent estimators of the PATE (e.g. Buchanan et al., 2018; Hartman et al., 2015; Stuart et al., 2011). In particular, recent papers explicitly show that researchers have to jointly consider treatment effect heterogeneity and the sampling mechanism (Cole & Stuart, 2010; Kern et al., 2016; Pearl & Bareinboim, 2014; Tipton, 2013). These existing approaches often assume researchers have access to a large number of covariates in both the experimental sample and the non-experimental target population. In contrast, we provide a new data-driven covariate selection algorithm to find separating sets in situations where researchers have data constraints in the target population. Our focus on covariate selection is similar to recent influential work on causal directed acyclic graphs (causal DAGs) (Bareinboim & Pearl, 2016; Bareinboim et al., 2014). We differ from the DAG-based approaches in that we empirically estimate separating sets under assumptions about sampling and heterogeneity sets rather than analytically selecting separating sets from fully specified causal DAGs. Although assumptions about the entire causal DAGs are sufficient for covariate selection, the proposed algorithm can estimate separating sets under weaker assumptions about sampling and heterogeneity sets at the expense of statistical uncertainty.

Research in the second direction argues that the necessary assumptions for existing methods are often too strong in practice. Recent papers have explored methods for sensitivity analyses (e.g. Andrews & Oster, 2019) and bounds (e.g. Chan, 2017) to achieve partial identification under weaker assumptions. In particular, Nguyen et al. (2017, 2018) also consider data scenarios where the experimental data have a richer set of covariates than the target population data, and propose a sensitivity analysis by specifying sensitivity parameters that captures the distribution of covariates unmeasured in the target population data. Our paper is complementary to these approaches. We instead focus on point identification of the PATE and alleviate strong assumptions about data requirements by adding an additional step of estimating a separating set. Our approach can also be used in conjunction with sensitivity analysis; the proposed method can estimate a smaller separating set, thereby reducing the number of unobserved covariates that sensitivity analyses have to consider. We provide further discussion about relationships between our proposed approach and sensitivity analysis in Section 4.4.

2 | YOUTH OPPORTUNITIES PROGRAM IN UGANDA

As well documented by the World Bank, a large number of young adults in developing countries are unemployed or underemployed (World Bank, 2012). In addition to its direct implication to poverty, concerns for policy makers are that such large young and unemployed populations can increase risk of crime and social unrest (Blattman et al., 2013). Uganda, especially conflict-affected Northern Uganda, is not an exception. According to estimates from the government, two-thirds of northern Ugandans could not meet basic needs, about 50% were illiterate, and most were underemployed in subsistence agriculture in 2006 (Government of Uganda, 2007).

In this paper, we study the Youth Opportunities Program (YOP) in Uganda, designed to help the poor and unemployed become self-employed artisans and increase incomes. This intervention is one example of widely used cash transfer programs in which participants are offered a certain amount of cash in the hope that they invest in training and start new, profitable enterprises. In 2008, the government invited young adults in Northern Uganda to form groups and submit grant proposals for how they would use a grant for vocational training and business start-up. Then, funding was randomly assigned among 535 screened, eligible applicant groups—265 and 270 groups to treatment and control respectively. Treatment groups received a one-time unsupervised grant worth $7,500 on average—about $382 per group member, roughly their average annual income. Following the original analysis, we focus on a binary treatment, whether they receive any grants or not through the YOP.

To evaluate the impact of this intervention, Blattman et al. (2013) surveyed five people per group three times over 4 years, resulting in a panel of 2,598 individuals after removing 79 observations due to missing data. They measured 17 outcome variables across five dimensions—employment (7), income (2), investments (3), business formality (3) and urbanization (2). They find that the effects of the YOP are large across all dimensions. Notably, after 2 years, the treatment groups were 4.5 times more likely to have vocational training, 2.6 times more likely to engage with a skilled trade and had 16% more hours of employment and 42% higher earnings.

Although it is unambiguous that the YOP had large, persistent positive effects on experimental subjects, it is of great policy interest to empirically investigate how much these experimental estimates are generalizable to a larger population. Estimating PATEs can inform which specific development policies governments should scale up. While the focus of the program was on Northern Uganda as a whole, participants of the YOP were inevitably not representative, as

TABLE 1 Estimates of sample average treatment effects and population average treatment effects based on the original eight variables

	SATE estimate	Original PATE estimate		SATE estimate	Original PATE estimate
Employment			**Investments**		
Average employment hours	5.13	5.87	Vocational training	0.55	0.51
	(1.22)	(3.25)		(0.03)	(0.07)
Agricultural	−0.01	0.75	Hours of vocational training	349.65	277.18
	(1.02)	(2.01)		(23.61)	(51.21)
Nonagricultural	5.14	5.12	Business assets	426.17	400.31
	(0.92)	(2.62)		(82.79)	(125.19)
Skilled trades only	4.75	2.73			
	(0.64)	(1.56)	**Business Formality**		
No employment hours	−0.02	0.00	Maintain records	0.12	0.19
	(0.01)	(0.02)		(0.03)	(0.07)
Any skilled trade	0.27	0.24	Registered	0.06	0.07
	(0.03)	(0.07)		(0.02)	(0.04)
Works mostly in a skilled trade	0.06	−0.01	Pays taxes	0.08	−0.01
	(0.01)	(0.03)		(0.02)	(0.06)
Income			**Urbanization**		
Cash earnings	13.41	12.15	Changed parish	0.01	−0.03
	(3.87)	(5.30)		(0.02)	(0.07)
Durable assets	0.11	0.09	Lives in Urban area	−0.01	−0.06
	(0.05)	(0.16)		(0.03)	(0.07)

Note: We estimated population average treatment effects (PATE) of the above 17 outcomes using an inverse probability weighting estimator with standard errors clustered by group. Weights are estimated by a logistic regression including the eight variables additively. See details of the estimation in Section 5. As a reference, we also report estimates of the sample average treatment effect (SATE)

in many other development programs. To take into account differences between experimental samples and Northern Uganda's population, Blattman et al. (2013) merged their experimental samples with a 2008 population-based household survey, the Northern Uganda Survey (NUS). They adjusted for eight variables shared by experimental and population data; gender, age, urban status, marital status, school attainment, household size, durable assets and district indicators.

In Table 1, we report estimates based on an inverse probability weighting (IPW) estimator (Stuart et al., 2011) that adjusts for the original eight variables.[1] As a reference, we also report estimates of the average treatment effect within the experimental sample, called the sample average treatment effect (SATE). Estimates of the SATEs and PATEs have roughly the same sign,

[1]Although the original authors rely on weighted linear regression models in their paper, we focus on the IPW estimator widely studied in the literature of generalization (e.g. Buchanan et al., 2018).

suggesting that the program will have a positive impact on a variety of outcomes even in the target population. Importantly, this finding, however, rests on an assumption that the original eight variables adjust for all relevant differences between the experimental sample and the target population. Given that the magnitude of the PATE estimates has strong implications for a cost-benefit analysis of these large-scale expensive interventions, it is critical to examine this common methodological challenge of covariate selection for generalizing experimental results.

In practice, there are several pervasive concerns about covariate selection. First, although it is common to adjust for all observed covariates shared by experimental and population data, it is unclear whether such sets of covariates include all necessary covariates for generalization. In fact, the authors carefully pay attention to this point in the original paper; 'young adults are selected into our sample because of unobserved initiative, connections or affinity for entrepreneurship' (Blattman et al., 2013). If there are unobserved differences between the experimental and population samples, the original PATE estimate would be biased. Second, it is also possible that the original analysis adjusted for unnecessary variables, resulting in inefficient estimators of the PATE. Miratrix et al. (2018) show that weighting on many variables, particularly those not highly correlated with treatment effect heterogeneity, can lead to inefficient estimation of the PATE. In this paper, we investigate necessary and sufficient sets of covariates for generalizing experimental estimates, called separating sets, and then provide a new algorithm to empirically estimate such sets. We select the separating sets under several different assumptions and assess how estimates of the PATE vary. Our reanalysis of this experiment appears in Section 5.

3 | SEPARATING SETS FOR GENERALIZATION

This section sets up the potential outcomes framework (Neyman, 1923; Rubin, 1974) for studying PATEs. We review a definition of *a separating set*—a set of variables affecting both the sampling mechanism and treatment effect heterogeneity, and then show that a sampling set and a heterogeneity set, the main focus of existing approaches, are special cases of the separating sets.

3.1 | The Setup

We consider a scenario in which we have two data sets. Following Buchanan et al. (2018), we define the first sample of n individuals to be participants in a randomized experiment ('Experimental Data') and the second data set to be a random sample of m individuals from the target population ('Population Data'). In our application, the experimental data have 2,598 individuals and the population data contain 21,348 individuals. We define a sampling indicator S_i taking 1 if unit i is in the experiment and 0 if unit i is in the target population. We assume that every unit has non-zero probability of being in the experiment.[2] Although experimental units can be randomly sampled from the target population in ideal settings, units often non-randomly

[2]This assumption of non-zero sampling probability is known as the 'Positivity of trial participation' (Colnet et al., 2020), and is commonly made in the generalization literature. This assumption is untestable and researchers have to evaluate it with domain knowledge (Dahabreh et al., 2020). When this assumption is violated, researchers have to rely on modelling assumptions and model-based extrapolation (Nethery et al., 2019). Alternatively, researchers can restrict their attention to a subset of the target population that has non-zero sampling probability (e.g. Tipton et al., 2016), which has been the most common approach in the causal inference literature to deal with non-overlap between the treatment and control groups.

select into the experiment, as in the YOP, making the experimental sample non-representative. Note that we consider cases in which units are either in the experimental data or in the target population data, but similar results hold for cases in which the experimental sample is a subset of the target population.

Let T_i be a binary treatment assignment variable for unit i with $T_i = 1$ for treatment and 0 for control. We define $Y_i(t)$ to be the potential outcome variable of unit i if the unit was to receive the treatment t for $t \in \{0,1\}$. In this paper, we make a stability assumption, which states that there is neither interference between units nor different versions of the treatment, either across units or settings (Hartman et al., 2015; Rubin, 1990; Tipton, 2013). We define pre-treatment covariates $\boldsymbol{X}_i$ to be any variables not affected by the treatment variable.

We are interested in estimating the average treatment effect in the target population. We call this causal estimand the PATE.

Definition 1 (Population Average Treatment Effect)

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0) \,|\, S_i = 0],$$

where $S_i = 0$ represents the target population data.

The treatment assignment mechanism is controlled by researchers within the experiment ($S_i = 1$), but it is unknown for units in the target population ($S_i = 0$; observational data). Formally, we assume that the treatment assignment is randomized within the experiment.

Assumption 1 (Randomization in Experiment)

$$\{Y_i(1), Y_i(0), \boldsymbol{X}_i\} \perp\!\!\!\perp T_i \,|\, S_i = 1$$

This assumption holds by design in randomized experiments. Here, we consider unconditional randomization, but results in the paper can be naturally extended to settings with randomization conditional on some pre-treatment covariates. Finally, for each unit in the experimental condition, only one of the potential outcome variables can be observed, and the realized outcome variable for unit i is denoted by $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (Rubin, 1974).

3.2 | Definition of Separating Sets and Identification

Recent papers show that the PATE can be identified by a set of variables affecting both treatment effect heterogeneity and the sampling mechanism (Cole & Stuart, 2010; Kern et al., 2016; Pearl & Bareinboim, 2014; Tipton, 2013). In this paper, we refer to this set as a *separating set* and investigate its statistical properties. Formally, a separating set is any set that makes the sampling indicator and treatment effect heterogeneity conditionally independent.

Definition 2 (Separating Set) A separating set is a set $\boldsymbol{W}$ that makes the sampling indicator and treatment effect heterogeneity conditionally independent.

$$Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \,|\, \boldsymbol{W}_i. \tag{1}$$

This definition of a separating set contains two simple cases: (1) when no treatment effect heterogeneity exists and (2) when the experimental sample is randomly drawn from the target population. In both of these cases, $\boldsymbol{W}_i = \{\varnothing\}$. This separating set also encompasses two common approaches in the literature as special cases. First, researchers often employ statistical methods based on a *sampling set*—a set of all variables affecting the sampling mechanism (e.g. Stuart et al., 2011). Second, researchers might adjust for a *heterogeneity set*—a set of all variables governing treatment effect heterogeneity (e.g. Kern et al., 2016). Below, we formalize these sets based on the potential outcomes framework.

We define a *sampling set* as a set of variables that determines the sampling mechanism by which individuals come to be in the experimental sample. For example, when a researcher implements stratified sampling based on gender and age, the sampling set consists of those two variables. When researchers control the sampling mechanism, a sampling set is known by design. However, when samples are selected without such an explicit sampling design, a sampling set is unknown and in practice, researchers must posit a sampling mechanism. For example, Blattman et al. (2013) assume that a sampling set consists of eight variables: gender, age, urban status, marital status, school attainment, household size, durable assets and district indicators.

Formally, we can define a sampling set $\boldsymbol{X}^S$ as follows.

Definition 3 (Sampling Set)

$$\{Y_i(1), Y_i(0), \boldsymbol{X}_i^{-S}\} \perp\!\!\!\perp S_i \,|\, \boldsymbol{X}_i^S \tag{2}$$

where $\boldsymbol{X}^{-S}$ is a set of pre-treatment variables that are not in $\boldsymbol{X}^S$.

This conditional independence means that the sampling set is a set that sufficiently explains the sampling mechanism. Given the sampling set, the sampling indicator is independent of the joint distribution of potential outcomes and all other pre-treatment covariates. We refer to variables in the sampling set as sampling variables.

The other popular approach is to adjust for a set of all variables explaining treatment effect heterogeneity, which we call a *heterogeneity set*. Formally, we can define a heterogeneity set $\boldsymbol{X}^H$ as follows.

Definition 4 (Heterogeneity Set)

$$Y_i(1) - Y_i(0) \perp\!\!\!\perp \{S_i, \boldsymbol{X}_i^{-H}\} \,|\, \boldsymbol{X}_i^H, \tag{3}$$

where $\boldsymbol{X}^{-H}$ is a set of pre-treatment variables that are not in $\boldsymbol{X}^H$.

In this case, because a heterogeneity set fully accounts for treatment heterogeneity, $Y_i(1) - Y_i(0)$ is independent of all other variables. We refer to variables in the heterogeneity set as heterogeneity variables. In our application, Blattman et al. (2013) discuss at least two heterogeneity variables, gender and initial credit constraints.

We want to emphasize that a sampling set and a heterogeneity set are special cases of a separating set in the sense that both sets satisfy Equation (1). Yet, there may exist many other separating sets, which we explore in Section 4.

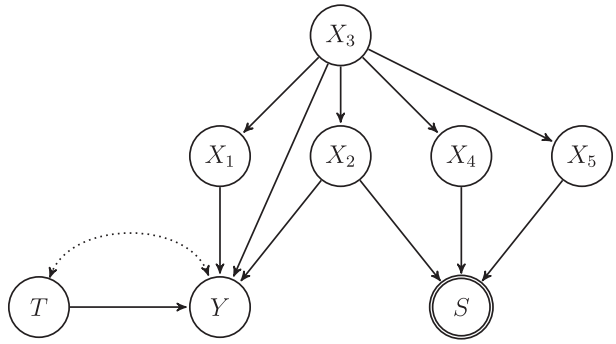


FIGURE 1 Example of a causal DAG based on a selection diagram approach (Bareinboim & Pearl, 2016). *Note:* Three variables $\{X_2, X_4, X_5\}$ serve as a sampling set, three variables $\{X_1, X_2, X_3\}$ as a heterogeneity set, and two variables as the smallest separating set $\{X_2, X_3\}$

Finally, the PATE is nonparametrically identified by adjusting for a separating set (Cole & Stuart, 2010; Kern et al., 2016; Pearl & Bareinboim, 2014; Tipton, 2013).

Result 1 (Identification of the PATE) The PATE is identified with separating set $\boldsymbol{W}_i$ under Assumption 1.

$$\tau = \int \left\{ \mathbb{E}[Y_i | T_i = 1, S_i = 1, \boldsymbol{W}_i = \boldsymbol{w}] - \mathbb{E}[Y_i | T_i = 0, S_i = 1, \boldsymbol{W}_i = \boldsymbol{w}] \right\} dF_{\boldsymbol{W}_i | S_i = 0}(\boldsymbol{w}),$$

where $F_{\boldsymbol{W}_i | S_i = 0}(\boldsymbol{w})$ is the cumulative distribution function of $\boldsymbol{W}$ conditional on $S_i = 0$.

As sampling and heterogeneity sets are special cases of a separating set, the PATE is identified with the same formula.

3.2.1 | Illustration with A Causal DAG.

We consider a causal DAG in Figure 1 as a concrete illustration based on a selection diagram approach (Bareinboim & Pearl, 2016). In this causal DAG, three variables $\{X_2, X_4, X_5\}$ serve as a sampling set, which has direct arrows to sampling variable S. Three variables $\{X_1, X_2, X_3\}$ serve as a heterogeneity set, which has direct arrows to outcome variable Y and can moderate the causal effect of T on Y. Finally, from Definition 2, there are many valid separating sets, including the sampling and heterogeneity sets, but the smallest separating set is $\{X_2, X_3\}$, which makes the sampling indicator S and the outcome variable Y conditionally independent. The key is that there are potentially many valid separating sets, and researchers can use any of them for identification given their data constraint.

We only use the causal DAG in Figure 1 as an illustrative example, and our proposed approach relies only on assumptions we clarify in Section 4 using the potential outcomes. We do not assume knowledge of the underlying causal DAG structure. When knowledge of the underlying causal DAG is available, results in the causal diagram literature are of great importance, and we refer readers to Bareinboim and Pearl (2016).

4 | IDENTIFICATION AND ESTIMATION OF SEPARATING SETS

In this section, we first show that a variant of *separating* sets, which is sufficient for the identification of the PATE, is estimable even when a *sampling* set is unobserved in the population data as far as it is observed in the experimental data (Section 4.1). This is in contrast to existing methodologies which assume that a sampling set is observed in *both* of the experimental and population data. This distinction is subtle and yet practically important because in many applied contexts, including the YOP (Blattman et al., 2013), a larger number of covariates are measured in the experimental data than in the population data. This is because, while analysts of experiments can often control what variables should be measured within the experiment, population data are usually more expensive; collected by other organizations, such as a national-level survey (the NUS in Blattman et al. (2013)); or otherwise impractical to collect. Thus, our focus is on this type of common research setting where analysts are able to measure more covariates in the experimental data than in the population data.

After demonstrating the identification of separating sets, we propose an algorithm to estimate separating sets using MRFs (Section 4.2).

4.1 | Identification of Separating Sets

We begin with the identification of an exact separating set and then turn to the identification of a modified separating set under a weaker assumption.

First, we can estimate an exact separating set in settings where both a sampling set and a heterogeneity set are observed in the experimental data. A key feature of this result is that we only require rich covariate information about the experimental units, not the target population units, to discover separating sets, should they exist.

In many applied research contexts, however, the heterogeneity set is not readily available even in the experimental data. The fundamental problem of causal inference states that only one of two potential outcomes are observable, which implies that the causal effect is unobserved at the unit level, and thus so is the heterogeneity set. For example, in our application, although Blattman et al. (2013) discuss two specific heterogeneity variables (gender and initial credit constraints), it might be unreasonable to assume away the existence of other potential heterogeneity variables.

We therefore develop an additional method to find a variant of a separating set, which we call a *marginal* separating set, using only knowledge of a sampling set, a commonly employed assumption in the extant literature. We show that a marginal separating set can be discovered when a sampling set is measured in the experimental data, but not in the target population. Although this data requirement might still be stringent in some contexts, it is much weaker than the one necessary for widely used existing approaches based on sampling sets, which require that sampling set be measured in the population data as well as in the experimental data.

4.1.1 | Identification of Exact Separating Sets

We begin with settings in which a sampling set and a heterogeneity set are observed in the experimental sample. In this setting, we can use the experimental data to identify exact separating

sets. Although this data requirement is still restrictive, we emphasize that it does not require rich data on the target population.

Setting 1 (Sampling and Heterogeneity Sets are Observed in Experiment) Sampling set $\boldsymbol{X}^S$ and heterogeneity set $\boldsymbol{X}^H$ are observed in the experiment ($S_i = 1$).

In this setting, a separating set is estimable as a set that makes the sampling set and the heterogeneity set conditionally independent within the experimental data.

Theorem 1 *(Identification of Separating Sets in Experiment) In Setting 1, consider a set of covariates $\boldsymbol{W}$, which is a subset of pre-treatment variables. Under Assumption 1,*

$$\tilde{\boldsymbol{X}}_i^H \perp\!\!\!\perp \tilde{\boldsymbol{X}}_i^S \,|\, \boldsymbol{W}_i, T_i, S_i = 1 \Longrightarrow Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \,|\, \boldsymbol{W}_i, \tag{4}$$

where $\tilde{\boldsymbol{X}}^H$ and $\tilde{\boldsymbol{X}}^S$ are the set difference $\boldsymbol{X}^H \setminus \boldsymbol{W}$ and $\boldsymbol{X}^S \setminus \boldsymbol{W}$ respectively.

We provide the proof in the supplementary material. Theorem 1 states that as long as we can find a set that satisfies the testable conditional independence on the left-hand side, the discovered set is guaranteed to be a separating set. That is, we can identify an exact separating set from the experimental data alone. Note that when $\boldsymbol{X}^H$ and $\boldsymbol{X}^S$ share some variables, those variables should always be in $\boldsymbol{W}$. Using the selected separating set, researchers can identify the PATE based on Result 1.

Intuitively, this theorem can be explained through two conceptual steps. First, because a heterogeneity set $\boldsymbol{X}^H$ fully explains treatment effect heterogeneity $Y(1)-Y(0)$, the sampling indicator S and $Y(1)-Y(0)$ are conditionally dependent only when S and $\boldsymbol{X}^H$ are conditionally dependent. Second, because a sampling set $\boldsymbol{X}^S$ fully explains the sampling indicator S, S and $\boldsymbol{X}^H$ are conditionally dependent only when $\boldsymbol{X}^S$ and $\boldsymbol{X}^H$ are conditionally dependent. Taken together, S and $Y(1)-Y(0)$ are conditionally dependent only when $\boldsymbol{X}^S$ and $\boldsymbol{X}^H$ are conditionally dependent.

Example. Here, we illustrate the general result from Theorem 1 using the causal DAG in Figure 1 as a concrete example. Suppose we are interested in a separating set $\boldsymbol{W} = \{X_2, X_3\}$, which has the smallest size. Given that heterogeneity set $\boldsymbol{X}^H = \{X_1, X_2, X_3\}$ and $\boldsymbol{X}^S = \{X_2, X_4, X_5\}$, we have $\tilde{\boldsymbol{X}}^H = X_1$ and $\tilde{\boldsymbol{X}}^S = \{X_4, X_5\}$. Then, as Equation (4) suggests, we have $X_1 \perp\!\!\!\perp \{X_4, X_5\} \,|\, X_2, X_3, T, S = 1$ in the causal DAG in Figure 1. More generally, Theorem 1 guarantees that any set that makes the sampling set and the heterogeneity set conditionally independent within the experimental data is a separating set under Assumption 1.

4.1.2 | Identification of Marginal Separating Sets

While Theorem 1 allows us to discover separating sets using the experimental data, a key challenge would be to measure both a sampling set and a heterogeneity set in the experimental data. In particular, it is often difficult to measure the heterogeneity set in practice. We show that a modified version of a separating set—a *marginal* separating set—is estimable from the experimental data under a weaker assumption. We define a marginal separating set as follows.

Definition 5 (Marginal Separating Set) A marginal separating set is a set $\boldsymbol{W}$ that makes the sampling indicator and the marginal distributions of potential outcomes conditionally independent.

$$Y_i(t) \perp\!\!\!\perp S_i \,|\, \boldsymbol{W}_i \text{ for } t = \{0, 1\}. \tag{5}$$

We refer to this as a *marginal* separating set since it renders the marginal, not the joint, distribution of potential outcomes conditionally independent of the sampling process.

Now we turn to our final setting researchers may find themselves in—that the sampling set is observed only in the experimental data. Previous work using the sampling set assumes it is measured in both the experimental sample and the target population (e.g. Buchanan et al., 2018; Cole & Stuart, 2010; Hartman et al., 2015; Tipton, 2013). Since researchers often have much more control over what data are collected in the experiment, this final setting greatly relaxes the data requirements of the previous literature.

Setting 2 (Sampling Set is Observed in Experiment) Sampling set $\boldsymbol{X}^S$ is observed in the experimental data ($S_i = 1$).

Theorem 2 *(Identification of Marginal Separating Sets in Experiment) In Setting 2, consider a set of covariates $\boldsymbol{W}$, which is a subset of pre-treatment variables. Under Assumption 1,*

$$Y_i \perp\!\!\!\perp \boldsymbol{X}_i^S \,|\, \boldsymbol{W}_i, T_i, S_i = 1 \Longrightarrow Y_i(t) \perp\!\!\!\perp S_i \,|\, \boldsymbol{W}_i. \tag{6}$$

We provide the proof in the supplementary material. Theorem 2 states that as long as we can find a set that makes the observed outcome Y conditionally independent of the sampling set within the experimental data, the discovered set is guaranteed to be a marginal separating set. With a large enough sample size, we can find a marginal separating set from the experimental data alone. Intuition behind this theorem is similar to the one used for Theorem 1. Because the sampling set $\boldsymbol{X}^S$ fully explains the sampling indicator S, if the sampling indicator S and the potential outcome $Y(t)$ are conditionally dependent, the sampling set $\boldsymbol{X}^S$ and the observed outcome Y are also conditionally dependent. The marginal separating set may be larger than an exact separating set, as it may include covariates that explain the marginal potential outcomes but not treatment effect heterogeneity. Once we have discovered a marginal separating set using the experimental data, we can identify the PATE with this discovered set.

Result 2 (Identification of the PATE with Marginal Separating Sets) When a marginal separating set $\boldsymbol{W}$ is observed both in the experimental sample and the target population, the PATE is identified with the marginal separating set $\boldsymbol{W}$ under Assumption 1.

$$\tau = \int \left\{ \mathbb{E}[Y_i | T_i = 1, S_i = 1, \boldsymbol{W}_i = \boldsymbol{w}] - \mathbb{E}[Y_i | T_i = 0, S_i = 1, \boldsymbol{W}_i = \boldsymbol{w}] \right\} dF_{\boldsymbol{W}_i | S_i = 0}(\boldsymbol{w}).$$

We omit the proof because it is straightforward from the one of Result 1.

TABLE 2 Identifying the PATE under different data requirements

	Data requirements	
Set to adjust for	**Experiment**	**Target population**
Sampling set	Sampling set	Sampling set
Heterogeneity set	Heterogeneity set	Heterogeneity set
Estimated separating set (Theorem 1 under Setting 1)	Sampling Set Heterogeneity Set	User Specified Constraints
Estimated marginal separating set (Theorem 2 under Setting 2)	Sampling set	User Specified Constraints

Note: Many previous approaches assume that a sampling set or a heterogeneity set is measured in both the experimental sample and the target population (the first two rows). Our proposed approaches relax data requirements for the target population by introducing an additional step of estimating separating sets.

Finally, in Table 2, we compare existing methods with two proposed approaches. The first two rows show two common existing approaches based on sampling and heterogeneity sets respectively. Although the identification of the PATE in those settings is straightforward, it requires rich covariate information from the target population data as well as from the experimental sample. Our approach relaxes data requirements for the target population by introducing an additional step of estimating separating sets. In Setting 1 where we observe both a sampling set and a heterogeneity set in the experimental sample, we can identify exact separating sets from the experimental data alone (Theorem 1). Setting 2 only requires observing a sampling set in the experimental sample and we can identify marginal separating sets (Theorem 2). In the next subsection, we introduce an algorithm that can estimate separating sets subject to user specified data constraints in the target population.

4.2 | Estimation of Separating Sets

Here, we propose an estimation algorithm to find a marginal separating set. As shown in Theorem 2, the goal is to find a set that makes a sampling set and observed outcomes conditionally independent within the experimental data. We show how to apply MRFs to encode conditional independence relationships among observed covariates and then select a separating set. A similar algorithm can be used for finding an exact separating set.

Our estimation algorithm consists of four simple steps. We provide a brief summary here and then describe each step in order. Step 1: specify all variables in sampling set $\boldsymbol{X}^S$ based on domain knowledge, some of which might not be measured in the population data. Step 2: using the experimental data alone, estimate a MRF over an outcome, a treatment, the sampling set and observed pre-treatment covariates. Step 3: enumerate all simple paths[3] from Y to $\boldsymbol{X}^S$ in the estimated Markov graph. Step 4: find sets that block all the simple paths from Y to $\boldsymbol{X}^S$ in the estimated Markov graph.

[3]A simple path is a path in a Markov graph that does not have repeating nodes.

4.2.1 | Estimating Markov Random Fields

Theorem 2 implies that we can find a marginal separating set by estimating a set of variables $\boldsymbol{W}$ that satisfies the conditional independence, $Y_i \perp\!\!\!\perp \boldsymbol{X}_i^S \mid \boldsymbol{W}_i, T_i, S = 1$. To estimate this set, we employ a MRF. MRFs are statistical models that encode the conditional independence structure over random variables via graph separation rules. For example, suppose there are three random variables A,B and C. Then, $A \perp\!\!\!\perp B \mid C$ if there is no path connecting A and B when node C is removed from the graph (i.e. node C *separates* nodes A and B), so-called the global Markov property (Lauritzen, 1996). We review basic properties of the MRF in the supplementary material (SM-3), and we refer readers to Lauritzen (1996) for its comprehensive discussion. Using the general theory of MRFs, the estimation of a separating set can be recast as the problem of finding a set of covariates separating outcome variable Y and a sampling set $\boldsymbol{X}^S$ in an estimated Markov graph. Therefore, we can find a separating set that satisfies the desired conditional independence as far as we can estimate the MRF over $\{Y, T, \boldsymbol{X}^S, \boldsymbol{X}_0\}$ within the experimental data where we define $\boldsymbol{X}_0$ to be all pre-treatment variables measured both in the experimental and population data. We define $\boldsymbol{Z} \equiv \{\boldsymbol{X}^S, \boldsymbol{X}_0\}$ to be pre-treatment covariates from which we select a separating set.

Note that MRFs are used here to estimate conditional independence relationships of observed covariates as an intermediate step of estimating separating sets. Importantly, they are not used to estimate the underlying causal DAGs. As emphasized earlier in the paper, while we used a causal diagram (Figure 1) as an illustration, we only rely on domain knowledge about sampling sets, and we are not taking the causal graphical approach. Such an approach is powerful when knowledge of the full structure of the underlying causal DAG is available, which we do not assume in this paper.

To estimate a MRF, we use a mixed graphical model (Haslbeck & Waldorp, 2020; Yang et al., 2015), which allows for both continuous and categorical variables. More concretely, we assume that each node can be modelled as the exponential family distribution using the remaining variables.

$$\Pr(G_r \mid G_{-r}) = \exp\left\{ \alpha_r G_r + \sum_{h \neq r} \theta_{r,h} G_r G_h + \varphi(G_r) - \Phi(G_{-r}) \right\}, \tag{7}$$

where G_{-r} is a set of all random variables in a Markov graph except for variable G_r, base measure $\varphi(G_r)$ is given by the chosen exponential family, and $\Phi(G_{-r})$ is the normalization constant. For example, for a Bernoulli distribution, the conditional distribution can be seen as a logistic regression model.

$$\Pr(G_r \mid G_{-r}) = \frac{\exp(\alpha_r + \sum_{h \neq r} \theta_{r,h} G_h)}{\exp(\alpha_r + \sum_{h \neq r} \theta_{r,h} G_h) + 1}. \tag{8}$$

In general, we model each node using a generalized linear model conditional on the remaining variables. Using this setup, we can estimate the structure of the MRF by estimating parameters $\{\theta_{r,h}\}_{h \neq r}$; $\theta_{r,h} \neq 0$ for variable G_h in the neighbours of variable G_r and $\theta_{r,h} = 0$ otherwise. We estimate each generalized linear model with ℓ_1 penalty to encourage sparsity (Meinshausen & Bühlmann, 2006). Finally, using the AND rule, an edge is estimated to exist between variables G_r and G_h when $\theta_{r,h} \neq 0$

and $\theta_{h,r} \neq 0$. Researchers can also use an alternative OR rule (an edge exists when $\theta_{r,h} \neq 0$ *or* $\theta_{h,r} \neq 0$) and obtain the same theoretical guarantee of graph recovery.

4.2.2 | Estimating Separating Sets

Given the estimated graphical model, we can enumerate many different separating sets. First, we focus on the estimation of a separating set of the smallest size because it often produces more stable weights and thus improves estimation accuracy. It is important to note that this separating set might not be the smallest with respect to the underlying unknown DAG because MRFs do not encode all conditional independence relationships between variables. It is the smallest size among all separating sets estimable from MRFs.

We estimate this separating set from pre-treatment covariates $\boldsymbol{Z}$ as an optimization problem. A separating set should block all simple paths between outcome Y and variables in the sampling set $\boldsymbol{X}^S$. Therefore, we first enumerate all simple paths between Y and $\boldsymbol{X}^S$ and then find a minimum set of variables that intersect all paths.

Define q to denote the number of variables in $\boldsymbol{Z}$. We then define $\boldsymbol{d}$ to be a q-dimensional decision vector with d_j taking 1 if we include the j th variable of $\boldsymbol{Z}$ into a separating set and taking 0 otherwise. We use $\boldsymbol{P}$ to store all simple paths from Y to each variable in $\boldsymbol{X}^S$ where each row is a q-dimensional vector and its j th element takes 1 if the path contains the j th variable. With this setup, the estimation of the separating set of the smallest size is equivalent to the following linear programming problem given the estimated graphical model.

$$\min_{\mathbf{d}} \sum_{j=1}^{q} d_j \quad \text{s.t., } \mathbf{Pd} \geq \mathbf{1}.$$

where $\mathbf{1}$ is a vector of ones. The constraints above ensure that all simple paths intersect with at least one variable in a selected separating set, and the objective function just counts the total number of variables to be included into a separating set. Therefore, by optimizing this problem, we can find a set of variables with the smallest size that is guaranteed to block all simple paths.

It is important to emphasize that the estimation of the MRF is subject to uncertainty as any other statistical methods. In our application, we incorporate uncertainties about set estimation through bootstrap. We also investigate accuracy of the proposed algorithm through simulation studies in the supplementary material. We find that estimators based on estimated separating sets often have similar standard errors to the ones based on the true sampling set. Although our approach introduces an additional estimation step of finding separating sets to relax data requirements, it does not suffer from substantial efficiency loss.

4.2.3 | Incorporating Users' Constraints

One advantage of our approach is that we can allow flexibility for researchers to explicitly specify variables that they cannot measure in the target population. This is important in practice because it is often the case that researchers can measure a large number of covariates in the experimental data but they can collect relatively few variables in the target population. We can easily adjust the previous optimization problem to account for this restriction. Define $\boldsymbol{u}$ to be a q-dimensional vector with u_j taking 1 if we want to exclude the j th variable of $\boldsymbol{Z}$ from a separating set and taking

0 otherwise. As we define $\boldsymbol{X}_0$ to be those variables observed in both the experimental sample and the target population, $\boldsymbol{u}$ will place constraints on those covariates in $\boldsymbol{X}^S$ that are unobservable. Then, the optimization problem above changes as follows.

$$\min_{\mathbf{d}} \quad \sum_{j=1}^{q} d_j \quad \text{s.t., } \mathbf{Pd} \geq \mathbf{1} \text{ and } \boldsymbol{u}^\top \mathbf{d} = 0$$

In practice, it is possible that there exists no separating set, subject to user constraints. In our example, a true separating set could include social connections, which are not measured in the Northern Uganda Survey (the population data). In this case, there is no feasible separating set and our algorithm finds no separating set.

4.3 | Estimation of the Population Average Treatment Effect

To estimate the PATE with estimated separating sets, we use an inverse probability weighting estimator. First, we estimate a probability of being in the experiment $\Pr(S_i = 1 \mid \boldsymbol{W}_i)$, for example, using a logistic regression (Stuart et al., 2011; Westreich et al., 2017) with adjustment for the actual target population size (Buchanan et al., 2018).[4] Following Buchanan et al. (2018), we stack the experimental data and the population data, and $S_i = 1$ ($S_i = 0$) indicates that unit i belongs to the experimental data (the population data). We can then estimate generalization weights as

$$\pi_i = \frac{1}{\Pr(S_i = 1 \mid \boldsymbol{W}_i)} \times \frac{\Pr(S_i = 0 \mid \boldsymbol{W}_i)}{\Pr(S_i = 0)}, \tag{9}$$

where a usual inverse probability is adjusted by $\Pr(S_i = 0 \mid \boldsymbol{W}_i)/\Pr(S_i = 0)$ because the PATE is defined only with the population data, that is, $\mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0]$. Finally, we compute the inverse probability weighting estimator (Stuart et al., 2011).

$$\widehat{\tau} \equiv \frac{\sum_{i;S_i=1} \pi_i p_i T_i Y_i}{\sum_{i;S_i=1} \pi_i p_i T_i} - \frac{\sum_{i;S_i=1} \pi_i (1-p_i)(1-T_i) Y_i}{\sum_{i;S_i=1} \pi_i (1-p_i)(1-T_i)}, \tag{10}$$

where $p_i \equiv \Pr(T_i = 1 \mid S_i = 1, \boldsymbol{W}_i)$ is known by the experimental design. We prove its consistency in the supplementary material.

Researchers can also employ an outcome-model-based estimator and a doubly robust estimator for the PATE (Dahabreh et al., 2019, 2020). To maintain the clear comparison with the original analysis that uses a weighting approach, we will focus on the inverse probability weighting estimator (Equation (10)) in Section 5, while we also report results from the other two estimators in the supplementary material.

[4]To account for the fact that the target population data are a random sample from the actual target population, we follow Scott and Wild (1986); Buchanan et al. (2018) to estimate a weighted logistic regression. We use weights 1 for the experimental data, and use weights $m/(N-n)$ for the target population data where the size of the actual target population $N = 10{,}000{,}000$ (population size in Northern Uganda), the size of the experimental data $n = 2{,}598$ and the size of the target population data $m = 21{,}348$.

4.4 | Relationship with Sensitivity Analysis

Here, we want to clarify the relationship between our proposed approach and sensitivity analyses for generalization (e.g. Andrews & Oster, 2019; Nguyen et al., 2017, 2018). In particular, our proposed approach can also be used to simplify sensitivity analysis. Sensitivity analysis in general uses some sensitivity parameters to quantify certain aspects of unobserved covariates. For example, Nguyen et al. (2017, 2018) propose a sensitivity parameter that captures the distribution of covariates unmeasured in the target population, and Andrews and Oster (2019) introduce a sensitivity parameter to quantify the predictive power of unobserved covariates relative to that of observed covariates. When there is a much richer set of covariates in the experimental data than the target population data—the common scenario and the main focus of our paper, analysts naturally need to specify a large number of sensitivity parameters to account for such potentially many covariates that are not measured in the target population data.

In such scenarios, it is well known that sensitivity parameters become more difficult to interpret, and analysts need to add more parametric assumptions (e.g. additivity) to handle many unobserved variables. Our proposed approach can be used to first find the smallest separating set within the experimental data. Then, researchers do not need to consider all covariates that are unmeasured in the target population data, and they can focus on a potentially much smaller set of covariates estimated as a separating set. Researchers can then use their preferred sensitivity analysis technique to deal with a much smaller set of covariates that are unmeasured in the target population data.

It is also important to clarify the scope of our approach. While we relax a stringent assumption that a sampling set is measured in both experimental and target population data, Theorem (2) still assumes that a sampling set is observed at least in the experimental data. If researchers are worried that a sampling set is not observed even in the experimental data (i.e. not observed in either the experimental or target population data), sensitivity analysis for completely unobserved covariates might be of greater importance.

5 | EMPIRICAL ANALYSIS

Applying the proposed method, we examine the YOP described in Section 2. Our focus is on a central methodological challenge of covariate selection. In the original analysis, the authors adjusted for all eight variables shared by the experimental and population data. However, as noted in the original paper, it is unknown whether the original eight variables are a separating set necessary for estimating PATEs. To tackle this pervasive concern, we employ the proposed approach and select a separating set under two different assumptions about a sampling set and a heterogeneity set.

First, we incorporate domain knowledge about a heterogeneity set, while we maintain the original assumption about a sampling set. As explained in Section 3, by combining substantive information about a sampling set and a heterogeneity set, we can find a separating set, which can be much smaller than each one of the two. Relying on this smaller separating set, we find that point estimates are similar to estimates based on the original sampling set, but standard errors of the proposed approach are smaller for 14 of 17 outcomes that the original analysis studied. Incorporating domain knowledge about a heterogeneity set can help us find a smaller set of variables sufficient for the PATE estimation, thereby improving efficiency.

Second, we relax the original assumption about a sampling set—the shared eight variables contain all relevant variables, and we allow for two additional unobserved variables. In the conventional approach based on a sampling set, researchers cannot estimate PATEs under this assumption. In contrast, the proposed approach estimated appropriate separating sets for 12 of 17 outcomes, and thus, we can estimate the PATE for those 12 outcomes even with the two additional unobserved sampling variables. At the same time, we reveal that estimated PATEs are sensitive to the original assumption about the sampling mechanism for the other five outcomes.

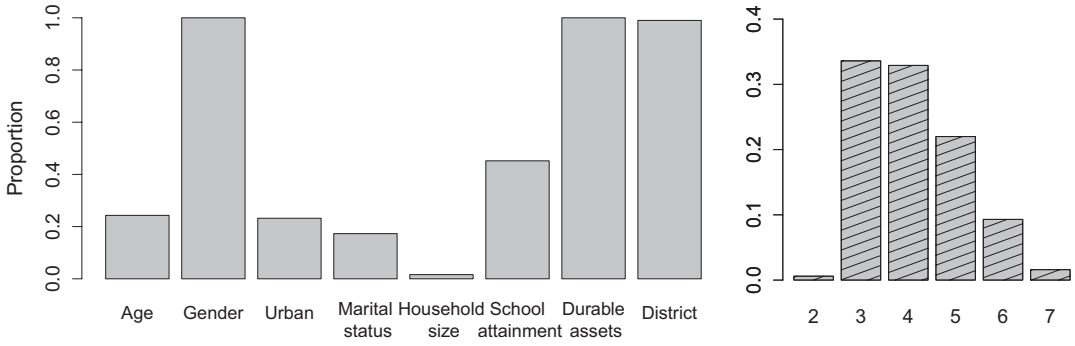
The original experiment used clustered randomization, and therefore, we compute clustered standard errors at the group level, which is the level at which treatments were randomly assigned. While we maintain the no interference assumption (Rubin, 1990) we made in Section 3.1, which is the standard assumption in the generalization literature, if researchers wish to account for interference within groups, it is important to additionally account for the difference in group structure between the experimental and target population data. Future work is necessary to consider this intersection of interference and generalization literature.

5.1 | Incorporating Domain Knowledge on Heterogeneity Set

To begin with, we maintain an assumption about a sampling set in the original analysis, that is, $\boldsymbol{X}^S = \{$`Gender, Age, Urban, Marital status, School attainment, Household size, Durable assets, District`$\}$. Although the original analysis relies only on this knowledge of the sampling set for the PATE estimation, the authors also carefully discuss a heterogeneity set in their paper. In particular, they discuss two variables: gender and initial credit constraints. There are two natural covariates in the experimental data that capture these concepts, `Gender` and `Initial Saving` respectively. Importantly, however, `Initial Saving` is measured only in the experimental sample and not in the target population data. Thus, $\boldsymbol{X}^H = \{$`Gender,` *Initial Saving*$\}$ where the italics represent a variable unmeasured in the target population.

In existing approaches, when a subset of heterogeneity variables are unmeasured in the target population as in this case, it is difficult to incorporate such domain knowledge into analysis and researchers often ignore the heterogeneity set altogether. In contrast, our proposed method uses knowledge about the heterogeneity set to estimate an exact separating set, which is potentially smaller than the observed sampling set and thus can increase the estimation accuracy for the PATE estimation. We first estimate a MRF over the union of sampling and heterogeneity sets within the experimental data. Then, we select an exact separating set that makes sampling set $\boldsymbol{X}^S$ and heterogeneity set $\boldsymbol{X}^H$ conditionally independent under a constraint that `Initial Saving` is unmeasured in the population data and cannot be selected. To take into account uncertainties, we estimate MRFs and select exact separating sets in each of 1000 bootstrap samples.

Figure 2 reports the results. The left panel (a) shows the proportion of each variable being estimated to be in an exact separating set over 1000 bootstrap samples. As the definition of separating sets (Definition 2) implies, the intersection of sampling and heterogeneity set, `Gender`, is always selected. In addition, `Durable assets` and `District` are selected almost always. Importantly, when we look at the size of estimated exact separating sets (the right panel (b)), it is often much smaller than the original size of eight (the mean size is 4.11). This means that even when a sampling set is sufficient for estimating the PATE, researchers can find a smaller separating set by incorporating domain knowledge of heterogeneity sets with the proposed approach.



(a) Proportion of Being Estimated as Separating Sets.

(b) Size of Estimated Separating Sets.

FIGURE 2 Estimated exact separating sets

Note: Panel (a) shows the proportion of each variable being estimated to be in an exact separating set over 1000 bootstrap samples. Panel (b) reports the size of estimated exact separating sets over 1000 bootstrap samples

If assumptions about the sampling set and the heterogeneity set hold, estimators based on the original sampling set $\boldsymbol{X}^S$ and on the estimated separating sets $\boldsymbol{W}$ are both consistent. However, standard errors of the latter might be smaller because corresponding estimated weights might be more stable.

To estimate the PATEs, we use the inverse probability weighting estimator proposed in Section 4.3. First, we estimate weights using the following logistic regression.

$$\text{logit}\{\Pr(S_i = 1 \mid \boldsymbol{C}_i)\} = \alpha_0 + \boldsymbol{C}_i^\top \beta, \tag{11}$$

where $\boldsymbol{C} = \boldsymbol{X}^S$ for the estimator based on the original sampling set and $\boldsymbol{C} = \boldsymbol{W}$ for our proposed estimator. We stack the experimental data (sample size = 2,598) and the population data (sample size = 21,348) and $S_i = 1(S_i = 0)$ indicates that unit i belongs to the experimental data (the population data). We can then estimate weights as $\widehat{\pi}_i = 1/\widehat{\Pr}(S_i = 1 \mid \boldsymbol{C}_i) \times \widehat{\Pr}(S_i = 0 \mid \boldsymbol{C}_i)/\widehat{\Pr}(S_i = 0)$, as proposed in Section 4.3. Note that treatment assignment probability in the experiment $\Pr(T_i = 1 \mid S_i = 1, \boldsymbol{W}_i)$ is equal to $\Pr(T_i = 1 \mid S_i = 1, \mathbf{D}_i)$ where $\mathbf{D}_i$ is a vector indicating 14 districts, because the treatment randomization was stratified by districts (Blattman et al., 2013). We use the block bootstrap to compute standard errors clustered at the group level as done in the original analysis. To take into account uncertainties of both steps—the estimation of separating sets and that of the PATE, in each of 1000 bootstrap samples, we estimate separating sets, estimate weights, and then estimate the PATE. Note that the difference between the estimator based on the original sampling set and our proposed estimator comes only from the selection of covariates $\boldsymbol{C}$ in the estimation of weights.

We report results in Table 3. Effects of the YOP are large and positive across many outcomes even among the broader target population. For example, the average employment hours would increase by 4.79 h (19% increase compared to the control group), monthly cash earnings would increase by 12,540 Uganda shilling (36% increase), and a proportion of people enrolled in vocational training would increase by 53 percentage points (349% increase). Comparing estimates based on the original sampling set and those based on the proposed separating set, we reveal that point estimates are similar to estimates with the original eight variables, and differences between them are not statistically significant at the conventional 0.05 level. This is expected because both

TABLE 3 Estimates of population average treatment effects based on the original set and the estimated separating set

	Original estimate	Sep. Set estimate		Original estimate	Sep. Set estimate
Employment			**Investments**		
Average employment hours	5.87	4.79	Vocational training	0.51	0.53
	(3.25)	(2.39)		(0.07)	(0.05)
Agricultural	0.75	0.30	Hours of vocational training	277.18	337.59
	(2.01)	(1.69)		(51.21)	(40.77)
Nonagricultural	5.12	4.49	Business assets	400.31	425.02
	(2.62)	(1.79)		(125.19)	(135.65)
Skilled trades only	2.73	4.36			
	(1.56)	(0.99)	**Business Formality**		
No employment hours	0.00	−0.03	Maintain records	0.19	0.20
	(0.02)	(0.03)		(0.07)	(0.07)
Any skilled trade	0.24	0.27	Registered	0.07	0.09
	(0.07)	(0.06)		(0.04)	(0.05)
Works mostly in a skilled trade	−0.01	0.04	Pays taxes	−0.01	0.05
	(0.03)	(0.03)		(0.06)	(0.05)
Income			**Urbanization**		
Cash earnings	12.15	12.54	Changed parish	−0.03	−0.01
	(5.30)	(5.11)		(0.07)	(0.04)
Durable assets	0.09	0.18	Lives in Urban area	−0.06	−0.01
	(0.16)	(0.13)		(0.07)	(0.04)

Note: We estimated population average treatment effects of 17 outcomes using weights based on the original eight variables ('Original estimate') and the estimated exact separating set ('Sep. Set estimate'). Standard errors of the proposed estimators are smaller for 14 of 17 outcomes.

estimators are consistent under the assumption that both specified sampling and heterogeneity sets are correct. More interestingly, we find that, for 14 of 17 outcomes, standard errors of estimators based on the estimated separating sets are smaller than those based on the original sampling set. On average, standard errors of the proposed approach are about 16% smaller. For the outcome 'Lives in Urban area', the standard error reduces more than 45%. This shows that by incorporating domain knowledge about heterogeneity sets, we can estimate smaller separating sets, which often improve efficiency.

5.2 | Accounting for Unobserved Sampling Set

In the previous analysis, we maintained the original authors' assumption about the sampling set and additionally took into account the assumption about the heterogeneity set. Here, we focus on estimating PATEs under weaker assumptions and directly address a concern noted in the original paper that the shared eight variable might not contain all relevant variables. In particular,

Blattman et al. (2013) discuss two potentially problematic variables. First, the authors are concerned that when the government screened applications at the village level, people with more social connections may have received some privilege. Second, people with 'affinity for entrepreneurship' (Blattman et al., 2013) might have been more likely to apply for the program in the first place. To account for these two sources of sample selection, we assume that a true sampling set contains two additional variables: (1) `Connection`, the number of community groups that a respondent belongs to, as a measure of social connections, and (2) `Business Advice`, total hours spent getting business advice in last 7 days, as a measure of initial motivation and affinity for entrepreneurship. Importantly, both of these two variables are not measured in the population data. Therefore, $\boldsymbol{X}^S = \{$`Gender, Age, Urban, Marital status, School attainment, Household size, Durable assets, District,` *Connection*, *Business Advice*$\}$ where the last two variables are measured only in the experiment and not in the population data. Moreover, we do not make any assumption about heterogeneity sets. Under this assumption, the current practice based on sampling sets or heterogeneity sets cannot estimate any PATEs; weights can be estimated only when sampling sets or heterogeneity sets are measured in both the experimental and population data. In contrast, the proposed method can select appropriate separating sets, should they exist, under such data constraints.

There are two questions of interest for each outcome; (1) Can we find a separating set and estimate the PATE? (2) If we can estimate the PATE, is an estimate different from the one based on the original eight variables? We estimate marginal separating sets using the proposed algorithm. For each outcome Y, we first estimate a MRF and then select a separating set that makes outcome Y and sampling set $\boldsymbol{X}^S$ conditionally independent under a constraint that the two unobserved variables (`Connection, Business Advice`) cannot be selected. When the algorithm can find no separating set under the constraint, we call it an 'infeasible solution'. We use block bootstrap to compute standard errors clustered at the group level as done in the original analysis. To take into account uncertainties over the covariate selection, we estimate MRFs and select separating sets in each of 1000 bootstrap samples.

We begin by computing proportions of infeasible solutions among the 1000 bootstraps (Figure 3). Proportions vary across outcomes, ranging from 0.0% ('Lives in Urban area') to 26.0% ('Registered'), and on average, 4.70%. Given that the current practice just based on sampling or heterogeneity sets cannot estimate PATEs for any outcomes, it is interesting that the proportions of infeasible solutions are smaller than 5% for 12 of 17 outcomes. For the remaining five outcomes, the average proportion of infeasible solutions is 11.5%, suggesting that the PATE estimates for these outcomes are sensitive to unobserved sampling variables, `Connection` and `Business Advice`.

For three outcomes that have less than 1% of infeasible solutions,[5] we also report estimates with 95% confidence intervals in Figure 4. We use the block bootstrap to compute standard errors clustered at the group level. To take into account uncertainties of both steps—the estimation of separating sets and that of the PATE, in each of 1000 bootstrap samples, we estimate separating sets, estimate weights, and then estimate the PATE. We find that point estimates are similar to the original estimates for the two outcomes ('Agricultural' and 'Lives in Urban area'), and for 'Changed parish', while point estimates are slightly different due to large standard errors, the difference between the original estimate and the estimate based on an estimated separating set is not statistically significant at the conventional level of 0.05. For all three outcomes, standard errors are smaller than the original ones. That is, estimates of the PATEs are robust to alternative separating sets, that is, even if the sampling set includes additional unobserved variables,

[5]'Agricultural' [0.6%], 'Changed parish' [0.9%], 'Lives in Urban area'[0.0%].

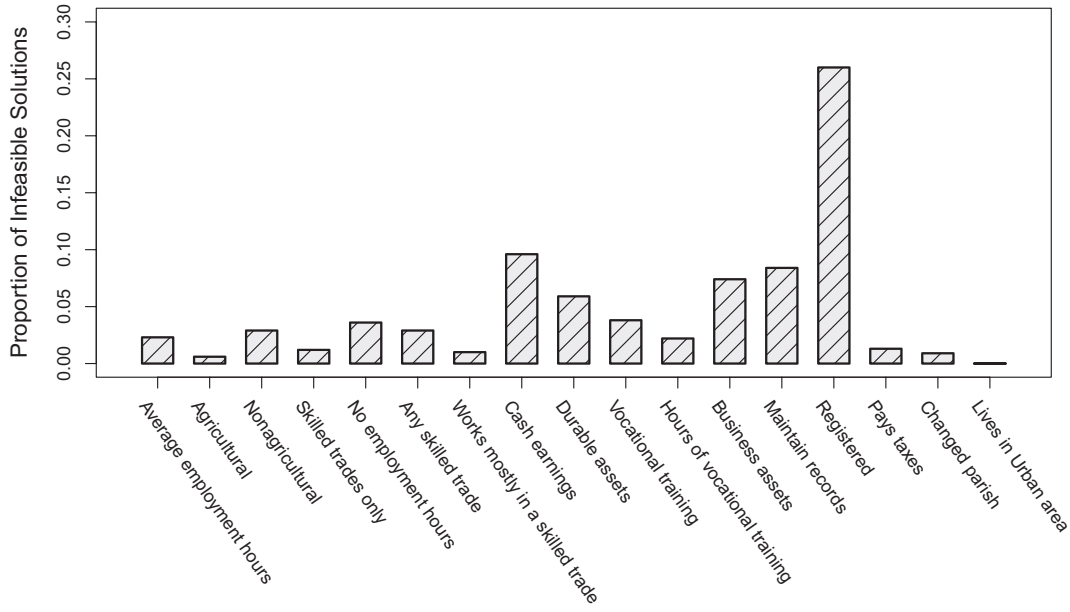


FIGURE 3 Proportions of infeasible solutions

Note: For 17 outcomes, we estimated marginal separating sets under a constraint that two sampling variables are unobserved in the population data. The figure shows the proportion of infeasible solutions for each outcome

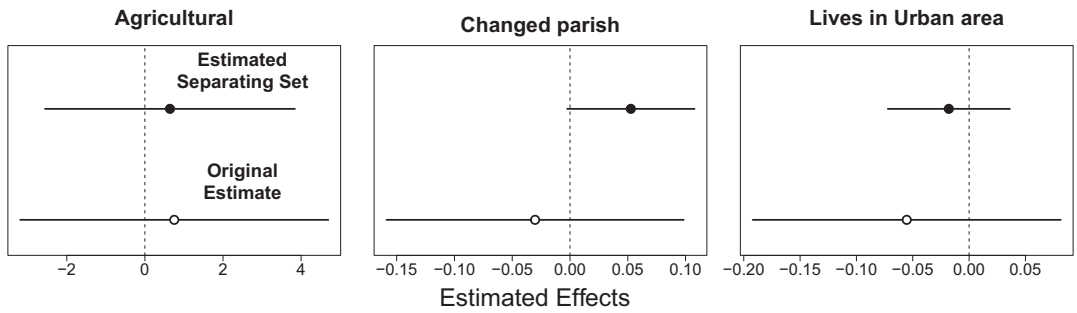


FIGURE 4 Estimates of population average treatment effects based on the original sets and estimated marginal separating sets

Note: We estimated population average treatment effects for three outcomes that have estimated proportions of infeasible solutions below 1%. Weights are based on the original eight variables ('Original') and estimated marginal separating sets ('Estimated Separating Set')

substantive conclusions are similar. This result demonstrates that the proposed algorithm of selecting separating sets allows researchers to estimate PATEs in situations where previous methods could not.

6 | CONCLUDING REMARKS

The increased emphasis on well-identified causal effects in the social and biomedical sciences can sometimes lead researchers to narrow the focus of their research question and limit their

findings to the experimental sample. However, primary research questions are often driven by the need to discover the impact of an intervention on a broader population. The extant literature has focused on the mathematical underpinnings concerning the generalizability of experimental evidence. The aim of this paper is to provide applied researchers with a means for uncovering a separating set using the experimental data alone.

Building on previous approaches, we clarify the role of the separating set—and its relationship to the sampling mechanism and treatment effect heterogeneity—in identification of PATEs. It makes clear that there are many possible covariate sets researchers can use for the recovery of population effects, and it allows us to develop a new algorithm that can incorporate researchers' data constraints on the target population.

As a concrete context, we focus on the YOP in Uganda. For these types of large-scale development programs, potential benefits and necessity of generalization are well known among researchers and policy makers. However, analysts are often constrained by available covariate information, which limits applicability of existing approaches that assume rich covariate data from both the experimental and population samples. Our proposed algorithm can help researchers to estimate appropriate separating sets, if any should exist, even under such data constraints. We find that by incorporating domain knowledge about heterogeneity sets, which is often overlooked in the PATE estimation, we can substantially improve efficiency. We also reveal that the proposed algorithm can find separating sets for 12 of 17 outcomes, even if we allow for two additional sampling variables that are not measured in the population.

Identifying population effects remains a challenging task for experimental researchers. The results here suggest researchers can increase a chance of generalization by collecting rich covariate information on their experimental subjects, even when their capacity of the population data collection is limited.

REFERENCES

Allcott, H. (2015) Site selection bias in program evaluation. *Quarterly Journal of Economics*, 130, 1117–1165.

Andrews, I. & Oster, E. (2019) A simple approximation for evaluating external validity bias. *Economics Letters*, 178, 58–62.

Angrist, J.D. & Pischke, J.-S. (2010) The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2), 3–30.

Bareinboim, E. & Pearl, J. (2016) Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27), 7345–7352.

Bareinboim, E., Tian, J. & Pearl, J. (2014) Recovering from Selection Bias in Causal and Statistical Inference. In Twenty-Eighth AAAI Conference on Artificial Intelligence.

Blattman, C., Fiala, N. & Martinez, S. (2013) Generating skilled self-employment in developing countries: Experimental evidence from Uganda. *The Quarterly Journal of Economics*, 129(2), 697–752.

Buchanan, A.L., Hudgens, M.G., Cole, S.R., Mollan, K.R., Sax, P.E., Daar, E.S., et al. (2018) Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 95, 1082.

Chan, W. (2017) Partially identified treatment effects for generalizability. *Journal of Research on Educational Effectiveness*, 10(3), 646–669.

Cole, S.R. & Stuart, E.A. (2010) Generalizing evidence from randomized clinical trials to target populations The ACTG 320 trial. *American Journal of Epidemiology*, 172(1), 107–115.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., et al. (2020) Causal inference methods for combining randomized trials and observational studies: a review.

Dahabreh, I.J., Robertson, S.E., Tchetgen, E.J., Stuart, E.A. & Hernán, M.A. (2019) Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2), 685–694.

Dahabreh, I.J., Robertson, S.E., Steingrimsson, J.A., Stuart, E.A. & Hernán, M.A. (2020) Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14), 1999–2014.

Deaton, A. & Cartwright, N. (2018) Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.

Druckman, J.N., Green, D.P., Kuklinski, J.H. & Lupia, A. (2011) *Cambridge handbook of experimental political science*. Cambridge: Cambridge University Press.

Duflo, E. & Kremer, M. (2005) Use of randomization in the evaluation of development effectiveness. In: Pitman, G., Feinstein, O. & Ingram, G. (Eds.) *Evaluating development effectiveness*, New Brunswick, NJ: Transaction Publishers, pp. 205–232.

Egami, N. & Hartman, E. (2020) Elements of External Validity: Framework, Design, and Analysis. Available from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id = 3775158.

Government of Uganda (2007) National Peace, Recovery and Development Plan for Northern Uganda: 2006–2009. Technical report, Government of Uganda, Kampala.

Hartman, E., Grieve, R., Ramsahai, R. & Sekhon, J.S. (2015) From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3), 757–778.

Haslbeck, J. & Waldorp, L.J. (2020) mgm: estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software*, 93(8), 1–46.

Imai, K., King, G. & Stuart, E.A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2), 481–502.

Imbens, G.W. (2010) Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2), 399–423.

Imbens, G.W. & Rubin, D.B.(2015) *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press.

Kern, H.L., Stuart, E.A., Hill, J. & Green, D.P. (2016) Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1), 103–127.

Lauritzen, S.L. (1996) *Graphical models*. Oxford: Clarendon Press.

Meinshausen, N. & Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3), 1436–1462.

Miratrix, L.W., Sekhon, J.S., Theodoridis, A.G. & Campos, L.F. (2018) Worth weighting? How to think about and use weights in survey experiments. *Political Analysis*, 26(3), 275–291.

Nethery, R.C., Mealli, F. & Dominici, F. (2019) Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The Annals of Applied Statistics*, 13(2), 1242.

Neyman, J. (1923) On the application of probability theory to agricultural experiments essay on principles (with discussion) section 9 (translated). *Statistical Science*, 5(4), 465–472.

Nguyen, T.Q., Ebnesajjad, C., Cole, S.R. & Stuart, E.A. (2017) Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, 11(1), 225–247.

Nguyen, T.Q., Ackerman, B., Schmid, I., Cole, S.R. & Stuart, E.A. (2018) Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, fffect scales, data scenarios, and implementation details. *PloS one*, 13(12), e0208795.

Pearl, J. & Bareinboim, E. (2014) External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4), 579–595.

Pressler, T.R. & Kaizar, E.E. (2013) The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Statistics in Medicine*, 32(20), 3552–3568.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.

Rubin, D.B. (1990) Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [Ann. Agric. Sci. 10 (1923), 1–51]. *Statistical Science*, 5(4), 472–480.

Scott, A.J. & Wild, C. (1986) Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2), 170–182.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stuart, E.A., Cole, S.R., Bradshaw, C.P. & Leaf, P.J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 369–386.

Stuart, E.A., Bradshaw, C.P. & Leaf, P.J. (2015) Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16(3), 475–485.

Tipton, E. (2013) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3), 239–266.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., et al. (2016) Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, 9(sup1), 209–228.

Westreich, D., Edwards, J.K., Lesko, C.R., Stuart, E. & Cole, S.R. (2017) Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186(8), 1010–1014.

World Bank (2012) World development report 2013: Jobs. Technical report, World Bank, Washington, DC.

Yang, E., Ravikumar, P., Allen, G.I. & Liu, Z. (2015) Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1), 3813–3847.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Egami N, Hartman E. Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda. *J R Stat Soc Series A*. 2021;184:1524–1548. https://doi.org/10.1111/rssa.12734